

Hacia un método computacional para detectar léxico contrastivo

Damián Eliel Aleman

30 de noviembre de 2017

Temario

1 Introducción

2 Trabajo Previo

3 Datos: Extracción y procesamiento

- Twitter
- Búsquedas geolocalizadas
- Tokenización y normalización

4 Métricas para detectar léxico contrastivo

- Primeras métricas: MaxDif y $MaxDif_g$
- Entropía y valor de la información
- Valores contrastivos

5 Análisis de las palabras contrastivas encontradas

- Caracterización de las palabras identificadas como contrastivas
- Validación estadística

6 Conclusiones y trabajo futuro

Motivación

- La palabra no está en el diccionario
- Conocer nuevas acepciones de palabras
- Diferenciar las marcas diatópicas de ciertas palabras

¿Qué podemos hacer al respecto?

Motivación

- La palabra no está en el diccionario
- Conocer nuevas acepciones de palabras
- Diferenciar las marcas diatópicas de ciertas palabras

¿Qué podemos hacer al respecto?

Método semi-supervisado para detectar léxico contrastivo

Qué es una palabra contrastiva

Se dice que una palabra es *contrastiva* cuando la frecuencia de uso en distintas regiones es muy diferente.

Ejemplos palabras contrastivas Argentina - España

- “che”
- “yapa”
- “tío”
- “metegol”
- “chabón”
- “chaval”

Ejemplos palabras contrastivas dentro de Argentina

- “gurisada”
- “chomaso”

¿Para qué sirve conocer las palabras contrastivas?

Temario

1 Introducción

2 Trabajo Previo

3 Datos: Extracción y procesamiento

- Twitter
- Búsquedas geolocalizadas
- Tokenización y normalización

4 Métricas para detectar léxico contrastivo

- Primeras métricas: MaxDif y $MaxDif_g$
- Entropía y valor de la información
- Valores contrastivos

5 Análisis de las palabras contrastivas encontradas

- Caracterización de las palabras identificadas como contrastivas
- Validación estadística

6 Conclusiones y trabajo futuro

Vidal de Battini

Separó a la Argentina en regiones dialectales



Método usual para detectar léxico contrastivo y las ventajas de complementar con la metodología computacional

¿Cómo se conocían las palabras contrastivas? **Mediante encuestas.**

- Costosas de realizar
- Muy difícil de hacer de forma balanceada en distintas regiones de un país/continente.
- Difícil de lograr un buen cubrimiento de hablantes
- Se basan en el conocimiento *a priori*

Temario

- 1 Introducción
- 2 Trabajo Previo
- 3 Datos: Extracción y procesamiento
 - Twitter
 - Búsquedas geolocalizadas
 - Tokenización y normalización
- 4 Métricas para detectar léxico contrastivo
 - Primeras métricas: MaxDif y $MaxDif_g$
 - Entropía y valor de la información
 - Valores contrastivos
- 5 Análisis de las palabras contrastivas encontradas
 - Caracterización de las palabras identificadas como contrastivas
 - Validación estadística
- 6 Conclusiones y trabajo futuro

Twitter

Qué es Twitter

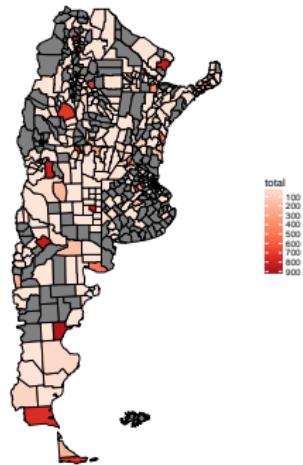
- Red social creada en el 2006
- Usuarios variados
- Tuits de hasta 140 caracteres^a
- **Todos los tuits son públicos**

^aRecientemente han aumentado el límite a 280 caracteres.

¿Por qué Twitter?

- API pública para obtener tuits de cualquier persona
- Tópicos muy variados (gran diferencia con portales de noticias, por ejemplo)
- Escalabilidad

Búsquedas geolocalizadas



- Se realizaron búsquedas en todos los departamentos de las provincias argentinas.
- Nos quedamos con los usuarios que ingresaron en el campo *location* al menos uno de los nombres de las ciudades de la provincia.

Distribución temporal de tuits

¿Y si en un momento dado la mayoría de los usuarios hablan del mismo tema por un fenómeno particular?

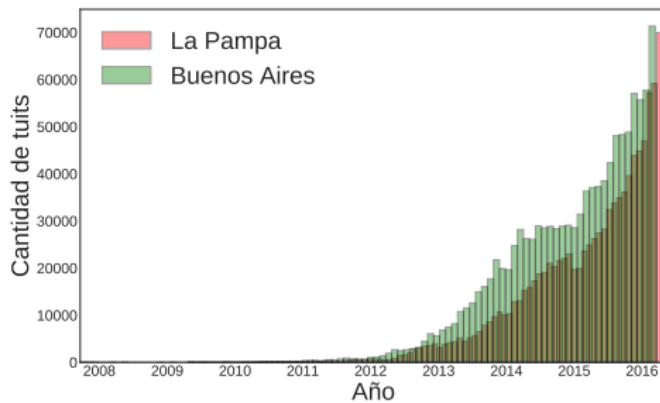


Figura: Histograma de la cantidad de tuits que se hicieron por intervalo de tiempo en la provincias La Pampa y Buenos Aires.

Balanceo de cantidad de tuits y de usuarios

Provincia	#Palabras Distintas	#Usuarios	#Tuits	#Total Palabras
Buenos Aires	191919	920	1125042	8974372
Catamarca	173104	957	1057019	8161309
Chaco	169476	964	976943	7605991
Chubut	182592	954	1023373	8884745
Córdoba	207307	987	1224266	10075932
Corrientes	183292	939	1044951	8426940
Entre Ríos	188679	969	1193693	9462986
Formosa	169254	903	923352	7184382
Jujuy	171064	971	678004	5951778
La Pampa	186593	935	1085757	8996318
La Rioja	186041	946	704044	6757277
Mendoza	193708	945	1099717	9402399
Misiones	168400	972	984218	7790197
Neuquén	188038	927	1111201	9021449
Río Negro	194383	965	1215361	9991831
Salta	188402	884	830916	7506652
San Juan	183546	926	1002322	8377792
San Luis	164185	896	1006464	8327093
Santa Cruz	174089	935	876621	7432923
Santa Fe	201879	937	1019620	8862328
S. del Estero	166540	887	944109	7355729
T. del Fuego	197273	964	976426	8559218
Tucumán	195643	962	1093874	9238526

Cuadro: Cantidades del conjunto de datos

¿Qué es una palabra?

Tokenización

Palabra = secuencia de caracteres alfabéticos (letras)

- #estalloElVerano → X
- ☺ ☺ → X
- https://nic.ar/ → X
- @BombitaDarin → X
- siempr3 → X
- siempre → ✓

Normalización

- **Minúscula:** Casa,CaSA,casA → casa
- **Hasta 3 repeticiones:** jajaaaaaaaa → jajaaaa

Cantidades por provincia

Palabra	Prov.1	Prov.2	...	Prov.23	Totales
w_1	$N_{1,1}$	$N_{1,2}$...	$N_{1,23}$	$\mathbf{N}(w_1)$
w_2	$N_{2,1}$	$N_{2,2}$...	$N_{2,23}$	$\mathbf{N}(w_2)$
.
w_i	$N_{i,1}$	$N_{i,2}$...	$N_{i,23}$	$\mathbf{N}(w_i)$
.
.
.

donde $N_{i,j}$ representa la cantidad de veces que la palabra w_i aparece en la provincia Prov. j , $j = 1, \dots, 23$.

$$\mathbf{N}(w_i) = \sum_{j=1}^{23} N_{i,j} \quad (1)$$

Temario

- 1 Introducción
- 2 Trabajo Previo
- 3 Datos: Extracción y procesamiento
 - Twitter
 - Búsquedas geolocalizadas
 - Tokenización y normalización
- 4 Métricas para detectar léxico contrastivo
 - Primeras métricas: MaxDif y $MaxDif_g$
 - Entropía y valor de la información
 - Valores contrastivos
- 5 Análisis de las palabras contrastivas encontradas
 - Caracterización de las palabras identificadas como contrastivas
 - Validación estadística
- 6 Conclusiones y trabajo futuro

Primeras métricas: MaxDif

Para cada palabra ω y cada par de provincias p_1 y p_2 :

$$maxDif(\omega, p_1, p_2) = \frac{\text{frec. máxima de } \omega \text{ en ambas provincias}}{\text{frec. mínima de } \omega \text{ en ambas provincias}} \quad (2)$$

Desventajas:

- ① Un valor para cada par de provincias.
- ② No se considera la dispersión de los valores en todas las provincias.

Primeras métricas: MaxDif

Para cada palabra ω y cada par de provincias p_1 y p_2 :

$$\maxDif(\omega, p_1, p_2) = \frac{\text{frec. máxima de } \omega \text{ en ambas provincias}}{\text{frec. mínima de } \omega \text{ en ambas provincias}} \quad (2)$$

\maxDif_g

Considerando las frecuencias de una palabra ω sobre todas las provincias:

$$\maxDif_g(\omega) = \frac{\text{frec. máxima de } \omega \text{ **todas las provincias**}}{\text{frec. mínima de } \omega \text{ sobre **todas las provincias**}} \quad (3)$$

- ① Se resume en un único valor la contrastividad de la palabra
- ② Sigue sin considerar la distribución de las frecuencias

La entropía de la información

Sea $\mathbf{p} = (p_1, \dots, p_n)$ un vector de probabilidad puntual: $p_i \geq 0$ y $\sum_{i=1}^n p_i = 1$.

Definimos la entropía de \mathbf{p} siendo

$$H(\mathbf{p}) = - \sum_{i=1}^n \log(p_i)p_i \quad (4)$$

La entropía de la información

Sea $\mathbf{p} = (p_1, \dots, p_n)$ un vector de probabilidad puntual: $p_i \geq 0$ y $\sum_{i=1}^n p_i = 1$.

Definimos la entropía de \mathbf{p} siendo

$$H(\mathbf{p}) = - \sum_{i=1}^n \log(p_i)p_i \quad (4)$$



Figura: Poco predecible,
alta entropía



Figura: Muy predecible,
baja entropía

La entropía de la información

Sea $\mathbf{p} = (p_1, \dots, p_n)$ un vector de probabilidad puntual: $p_i \geq 0$ y $\sum_{i=1}^n p_i = 1$.

Definimos la entropía de \mathbf{p} siendo

$$H(\mathbf{p}) = - \sum_{i=1}^n \log(p_i)p_i \quad (4)$$

Dime que tan uniforme eres y te diré cuanta entropía tienes

Las palabras utilizadas más uniformemente en las distintas provincias como *de* o *que* aportan menos información que la palabra *chomoso*.

La entropía de la información

Sea $\mathbf{p} = (p_1, \dots, p_n)$ un vector de probabilidad puntual: $p_i \geq 0$ y $\sum_{i=1}^n p_i = 1$.

Definimos la entropía de \mathbf{p} siendo

$$H(\mathbf{p}) = - \sum_{i=1}^n \log(p_i)p_i \quad (4)$$

Observaciones

- ① Entropía mínima ($= 0$) si y solo si \mathbf{p} esta concentrada en un único punto: existe i tal que $p_i = 1$ y $p_j = 0$, para todo $j \neq i$.
- ② Entropía máxima con \mathbf{p} equiprobable: $p_i = 1/n$, para todo i .

Valor de la información

Zanette y Montemurro definieron al *valor de la información de una palabra* como

$$\Delta I(\omega) = p(\omega) (\hat{H}(\omega) - H(\omega)) = p(\omega) \Delta H(\omega) \quad (5)$$

siendo $p(\omega)$ la frecuencia total de la palabra en el texto, $p(\omega) = n/N$
 $N = \#$ palabras en texto, $n = \#$ ocurrencias de ω

$$n_1 \dots n_p \rightarrow H(\omega) \quad (6)$$

Valor contrastivo sobre las palabras

Valor contrastivo sobre las palabras

$$I_w(\omega) = \text{norm}_w(\omega) \cdot (\hat{H}_w(\omega) - H_w(\omega)) \quad (7)$$

donde norm_w sirve para normalizar sobre la cantidad de ocurrencias de la palabra.

Observaciones

- Si una palabra se dice muchas veces el valor de $I_w(\omega)$ es más alto.
- Si dos palabras se dicen la misma cantidad de veces, la palabra que tenga una dispersión más heterogénea será la de mayor valor contrastivo sobre las palabras.

Robusteciendo métrica: valor contrastivo sobre las personas

¿Y si algunas palabras tienen un I_w alto debido a pocas personas que las mencionan constantemente?

Hay que tener en cuenta a la distribución de la cantidad de personas que mencionan cada palabra en las provincias.

Valor contrastivo sobre las personas

$$I_p(\omega) = \text{norm}_p(\omega) \cdot (\hat{H}_p(\omega) - H_p(\omega)) \quad (8)$$

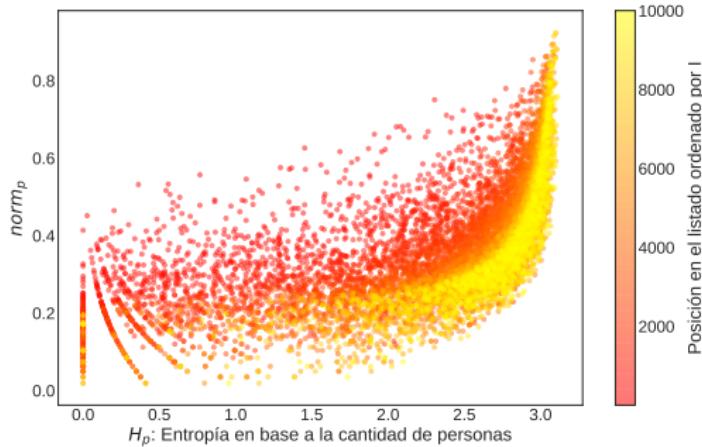
donde norm_p sirve para normalizar sobre la cantidad de personas que mencionan la palabra ω .

Robusteciendo métrica: valor contrastivo sobre las personas

Valor contrastivo sobre las personas

$$I_p(\omega) = \text{norm}_p(\omega) \cdot (\hat{H}_p(\omega) - H_p(\omega)) \quad (8)$$

donde norm_p sirve para normalizar sobre la cantidad de personas que mencionan la palabra ω .

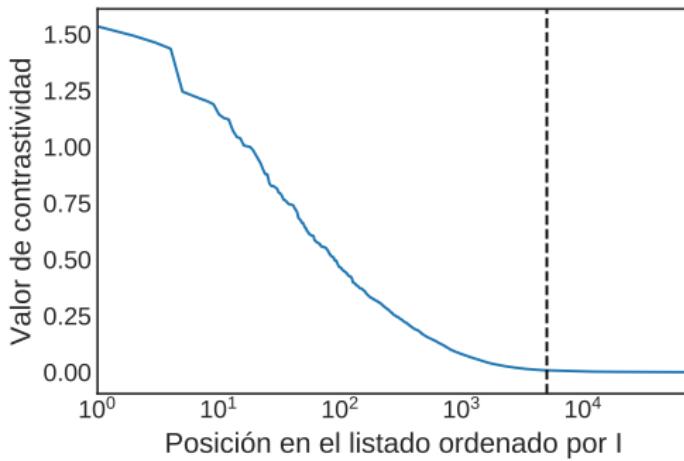


Valor de contrastividad

Valor de contrastividad

Como nos interesa tanto la distribución de la cantidad de ocurrencias de cada palabra, como la de la cantidad de usuarios que la menciona, definimos

$$I(\omega) = I_w(\omega) \cdot I_p(\omega) \quad (9)$$



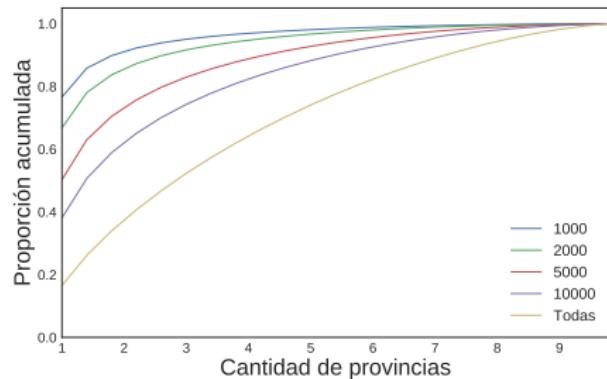
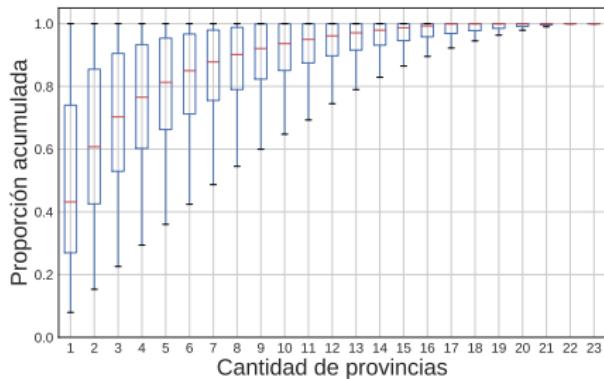
Proporción de ocurrencias

¿Dónde se utiliza más?

por cada palabra w :

por cada $k = 1..23$:

calculamos la proporción acumulada de ocurrencias
por las k provincias que más mencionan a w



Temario

- 1 Introducción
- 2 Trabajo Previo
- 3 Datos: Extracción y procesamiento
 - Twitter
 - Búsquedas geolocalizadas
 - Tokenización y normalización
- 4 Métricas para detectar léxico contrastivo
 - Primeras métricas: MaxDif y $MaxDif_g$
 - Entropía y valor de la información
 - Valores contrastivos
- 5 Análisis de las palabras contrastivas encontradas
 - Caracterización de las palabras identificadas como contrastivas
 - Validación estadística
- 6 Conclusiones y trabajo futuro

Pipeline



Validación lingüística

Trabajo realizado por la Academia Argentina de Letras (AAL):

- **Coloquialismos o vulgarismos**

“Q **chombi** hacer un chiste y q la otra persona no se ría o no lo entienda” (Mendoza)

“Que **carnasas** poniendole rosas rojas a toda la ropa, para mi queda horrible sorry” (Neuquén)

- **Indigenismos**

“Te regalo ser **mitaí** y ir a jurar la bandera con el guardapolvo caliente ese y la corbata que te ahorca todo (Del guaraní mitaí “pequeño”)” (Formosa)

“**Angá** mi negrito, esta triste (Del guaraní angá aprox. “pobre”)” (Corrientes)

Más resultados

- **Leísmo**

“No te olvides de **saludarle** a tu suegro hoy” (Misiones)

“Vine a **visitarte** a mis primas y estan re colgadas, para eso me quedaba en mi casa no maaa ” (Misiones)

“A **esperarle** a nahuel, que traiga los teressss ” (Formosa)

Más resultados

- **Voces sospechadas generales pero con acepción local diferente**
“**Mansas** ganas de sentarme a tomar un te con semitas” (San Juan)
“**sigo asada** por cosas que han pasado hace como dos dias, que falla (Mendoza) / Que **asada** estoy, tengo la cabeza echa un lío” (San Juan)

Más resultados

- **Voces con una morfología propia de una región**

Ejemplo: terminación azo/aza con base adjetiva.

“Esta **locaza** esa mina para hacer eso” (Córdoba)

Más resultados

- **Formas verbales coloquiales con sustantivos o adjetivos como base**

“Me calma mucho **mimosear** a mi perro ” (Neuquén)

“Estaría bueno que ari venga aunque sea a saludarme y que no se quede todo el tiempo **pollereando.**” (Tierra del Fuego)

Más resultados

- **Vesres:** Creación de palabras por inversión de sílabas que se usa jergalmente o con fines humorísticos.
“Estoy en lo de villa mateando con él y jimmy. Pinta **sogui** abundante más tarde dijeron ” (Corrientes)
“Uhhh me acuerdo si no habré saltado el muro del aguapey par colarme a los **cequin**. (cequín “fiesta de quince”)” (Chaco)

Más resultados

- **Intejercciones**

“**Aijué**, encima me decís vieja, re que no pinta esto facundo jaja ya te dije como es la onda, fin ” (Formosa)

“**Ains**, una mujer hablando de fútbol.” (Formosa)

“Al fin una buena: hora libreeee! **Yirr** ” (Corrientes)

Problema desde el punto de vista estadístico

Utilizar tests estadísticos para validar los resultados.

Hipótesis nula: H_0

La palabra tienen un uso homogéneo en las distintas regiones de la Argentina, es decir que la frecuencia de ocurrencias de cada palabra debería ser similar independientemente de la región.

Test t de Welch

El test de Welch nos provee un valor de probabilidad para rechazar la hipótesis nula: las medias de las dos distribuciones son iguales.

Las suposiciones del test

- ① Textos son estadísticamente independientes
- ② La media de las frecuencias proviene de una distribución normal

Metodología

Agrupamos todos los tuits de cada usuario representando un texto.

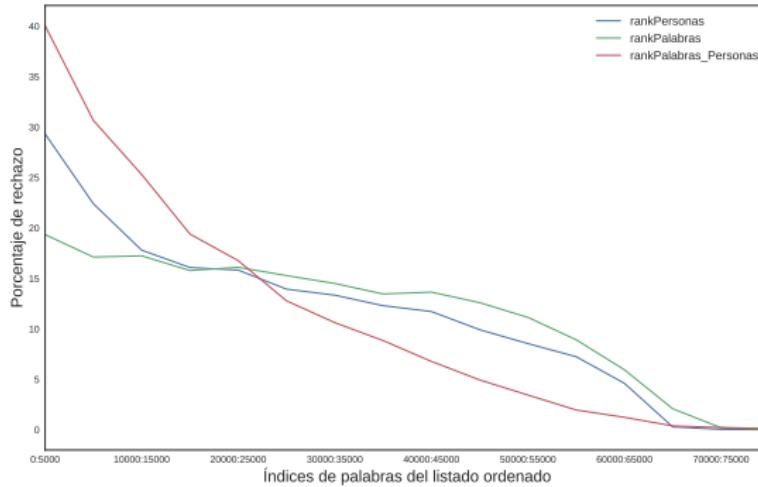
Corpus S Todos los textos sobre la región que cubre más del 80 % de las ocurrencias

Corpus T Los textos creados por usuarios del resto de las provincias

Resultados test t de Welch

- ① Para cada métrica I, I_W, I_P variamos los subconjuntos de palabras de acuerdo al listado ordenado según estas.
- ② Calculamos la tasa de rechazo de la hipótesis nula, definida por:

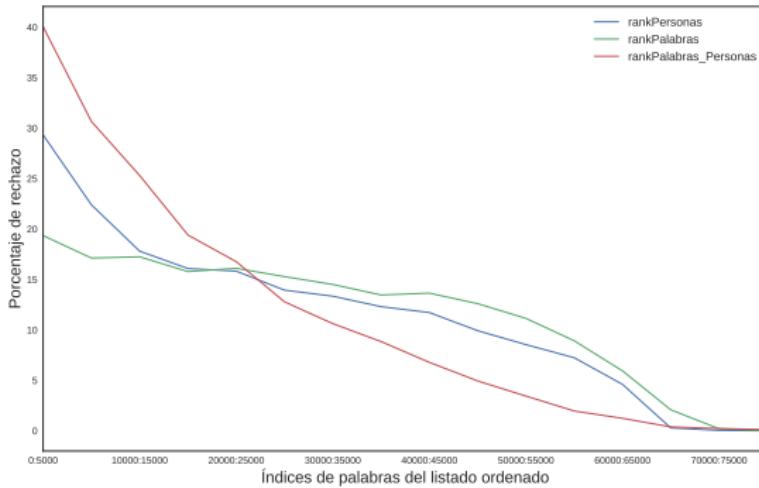
$$\text{Tasa de rechazo}(\text{tests}) = \frac{\#\{t : \text{tests} \mid p\text{-valor}(t) < 0,05\}}{\#\text{tests}} \quad (10)$$



Resultados test t de Welch

- Para cada métrica I, I_W, I_P variamos los subconjuntos de palabras de acuerdo al listado ordenado según estas.
- Calculamos la tasa de rechazo de la hipótesis nula, definida por:

$$\text{Tasa de rechazo(tests)} = \frac{\#\{t : \text{tests} \mid p\text{-valor}(t) < 0,05\}}{\#\text{tests}} \quad (10)$$



Temario

- 1 Introducción
- 2 Trabajo Previo
- 3 Datos: Extracción y procesamiento
 - Twitter
 - Búsquedas geolocalizadas
 - Tokenización y normalización
- 4 Métricas para detectar léxico contrastivo
 - Primeras métricas: MaxDif y $MaxDif_g$
 - Entropía y valor de la información
 - Valores contrastivos
- 5 Análisis de las palabras contrastivas encontradas
 - Caracterización de las palabras identificadas como contrastivas
 - Validación estadística
- 6 Conclusiones y trabajo futuro

Conclusiones

- Desarrollamos una métrica de la contrastividad de uso de una palabra en distintas regiones.
- Para probar esta métrica recolectamos un conjunto de datos de textos de la Argentina a través de la API de Twitter.
- Obtuvimos aproximadamente 300 palabras contrastivas relevantes lingüísticamente sobre las 5000 contrastivas.
- Varias de las palabras detectadas a partir de la métrica desarrollada serán agregadas al Diccionario del habla de los argentinos.

Trabajo a futuro

- Reproducir el trabajo para todos los países hispanoparlantes.
- Obtener regiones dialectales a partir de métodos de clustering, lo cual permitiría validar la vigencia de las regiones propuestas por Vidal de Battini en 1964 [VdB64].
- Analizar la contrastividad léxica comparando la distribución de n-gramas.

¿Preguntas?

Test hipergeométrico

Para aplicar el test hipergeométrico representamos los datos sobre la palabra en una tabla de 2x2 como la de la siguiente Tabla.

	#Palabras sobre región	#Palabras en el resto de Argentina	Total
# Palabras w	k	$K - k$	K
# Palabras $\neq w$	$n - k$	$N + k - n - K$	$N - K$
Total	n	$N - n$	N

$$cant_W(w) = \log(\mathbf{N}(w)), \quad (11)$$

y

$$MIN_W = \min_{w \in \text{Palabras}} cant_W(w), \quad MAX_W = \max_{w \in \text{Palabras}} cant_W(w) \quad (12)$$

Índice normalizado

$$norm_W(w) = \frac{cant_W(w) - MIN_W}{MAX_W - MIN_W} \quad (13)$$

Esto se define para palabras w con $w \geq 40$.

Volver



Berta Elena Vidal de Battini.

El español en la argentina.

Technical report, Argentina., 1964.